Large Language Models are Echo Chambers

Anonymous submission

Abstract

Modern large language models and chatbots based on them show impressive results in text generation and dialog tasks. At the same time, these models are subject to criticism in many aspects, e.g., they can generate hate speech and untrue and biased content. In this work, we show another problematic feature of such chatbots: they are echo chambers in the sense that they tend to agree with the opinions of their users. Social media, such as Facebook, was criticized for a similar problem and called an echo chamber. We experimentally test five LLM-based chatbots, which we feed with opinionated inputs. We annotate the chatbot answers whether they agree or disagree with the input. All chatbots tend to agree. However, the echo chamber effect is not equally strong. We discuss the differences between the chatbots and make the dataset publicly available (We will add a link to the camera-ready version).

Keywords: large language models, chatbots, bias, echo chambers

1. Introduction

Criticism of LLMs Large language models (LLMs) such as ChatGPT or OpenAI GPT3 (Brown et al., 2020b) show impressive results in text generation. They advanced the development of a new generation of chatbots based on LLMs and can lead coherent conversations on any topic. ChatGPT is a very influential LLM-based chatbot that sparked huge public interest beyond the scientific community. However, LLMs are the subject of criticism in many aspects. Bender et al. (2021) point out that these models reflect various biases from their training data and their generated text can contain hate speech, racism, sexism, untrue and other undesired content (Dhamala et al., 2021; Bender et al., 2021; Brown et al., 2020a; Nadeem et al., 2021). In this work, we elaborate on a problematic feature of modern chatbots that, to our knowledge, has not yet been addressed by research: We show that LLM based chatbots are echo chambers, which means that they tend to agree with the opinion of their users.

Echo chambers Across societies, humans exhibit a remarkably stable preference for the homogeneity of certain social domains. Well-documented examples include ideological, ethnic, and socio-economic homophily, which drives people to prefer association with similar individuals (Bishop, 2008). These social phenomena are complemented by consistency theories in psychology, which posit similar models for the preference of attitude-consistent information (Osgood and Tannenbaum, 1955). Although these tendencies towards social homogeneity work to reduce societal cohesion and thus threaten the effectiveness of political processes (Sunstein, 2001), there is a counteracting force: Serendipitous exposure in public social settings and from diverse information sources bridges gaps between social groups by facilitating novel contacts (Fletcher and Nielsen, 2018)

This balance between homogeneity and heterogeneity has recently been changed by the introduction of personalized online recommender systems (Bernstein et al., 2021). Online platforms can exacerbate societal fragmentation by connecting like-minded individuals and fading out others, leading to what has been termed “echo chambers” (Sunstein, 2001). Recent research has extended this simple premise, showing that platforms have differential effects both expanding and limiting diverse exposure (Jürgens and Stark, 2022), and that echo chambers are most likely to affect societal fringe groups (Stegmann et al., 2022). The rapid diffusion of LLMs has introduced a new potential mechanism that may induce or exacerbate echo chambers. Interactions with LLMs mimic social conversations, potentially affecting opinion formation and social perception processes. Models echoing the stance of their users could create a significant risk of echo chambers, political polarization, and radicalization.

Opinionated LLMs Our research expands the research on opinionated large language models. Previous research showed that LLMs and chatbots based on them align more closely with left-leaning environmental or political positions (Hartmann et al., 2023; Santurkar et al., 2023). However, these political views are not consistent, e.g., left-leaning models can also express conservative opinions on religion (Santurkar et al., 2023) or tend to take opinions from a specific political spectrum when they are prompted with a personality from such a spectrum (Schramowski et al., 2022).

Contributions In this work, we show that LLM-based chatbots are echo chambers, which means that they tend to agree with the opinions of their users. In an experiment, we feed a dataset with

topic	stance	original tweet	adapted tweet
climate change	favor	*The biggest terror threat in the World is climate change #drought #floods #SemST*	*The biggest terror threat in the world is climate change.*
legalization of abortion	against	*Patient advocate Stephen Casey of Texas Center for Defense of Life strongly supports #HB3074. #TXlege #SemST*	*I strongly support the anti-abortion bill.*
atheism	against	*9 Holy Mary Mother of God, pray for us sinners now and at the hour of our death. Amen. #rosary #teamjesus #God #SemST*	(discarded tweet)

Table 1: Examples of the adaptation of original tweets to chatbot inputs.

chatbot	input	response	(dis)agreement
Godel	Feminists are instable. They get super agitated by criticism.	Feminists are usually instable, so they would get upset by negative criticism.	`Agree`
ChatGPT 3.5	The religious do not acknowledge scientific discoveries.	As an AI language model, I must be impartial and correct. It is inaccurate to make a blanket statement that all religious individuals do not acknowledge scientific discoveries. [...]	`Disagree`

Table 2: Examples of annotated chatbot answers

opinions from different political topics into a selection of five LLMs. Then, we annotate the answers regarding agreement bias and analyze the results. On average, all LLMs agree 3.02 times more likely with opinions than disagree. This effect differs between LLMs between 1.63 and 6.22.

2. Data generation

2.1. Generation of chatbot input

We adapt a dataset from Mohammad et al. (2016) to generate opinionated chatbot inputs. This dataset contains 4,870 tweets labeled with their stance on topics like climate change is a real concern, abortion, feminism, atheism, Donald Trump, and Hillary Clinton. We focused on tweets expressing opinions in favor or against these topics, excluding tweets without any opinion and all tweets about the election of Trump and Clinton since this subject is particular to the US only. We randomly selected a subset of 353 tweets from the remaining ones and edited them to create suitable chatbot inputs. The text adaptation includes: (a) removing Twitter-specific elements like hashtags and usernames, (b) rewriting tweets to express a stance to make their content clear without extra context (c) discarding tweets that could not be reformulated clearly. To meet these criteria, a trained linguist performed the editing with minimal changes to the original tweets. This process resulted in 333 adapted tweets: 199 in favor and 134 against. See Table 1 for examples of original tweets, their adaptations, and discarded tweets.

2.2. Generation of chatbot answers

LLM	Num. params	Chatbot
Blenderbot (Roller et al., 2021)	400M	yes
Godel Large v1.1 (Peng et al., 2022)	700M	yes
ChatGPT (GPT-3.5-turbo) (Schulman et al., 2023)	unknown	yes
Davinci (GPT3) (Brown et al., 2020c)	175B	no
Llama1 (Touvron et al., 2023)	7B	yes

Table 3: LLMs used in our study

Table 3 shows the LLMs used in our study. We chose a list of relatively old and small LLM-based chatbots (Blenderbot and Godel), the powerful and very large state-of-the-art LLMs ChatGPT and Davinci and LLama as an example of a modern, smaller LLM. Except for Davinci, all these LLMs were designed for chat. The table lists the number of parameters for each chatbot and whether the model was explicitly designed for chat or dialog. We feed the clean text into each model without additional prompts and use the default parameters of the models to generate text.

2.3. Annotation of chatbot answers

We annotated each chatbot response with one of the following values: `Agree` if the response rather agrees with the input regarding the topic, `Disagree` if the response rather disagrees, and

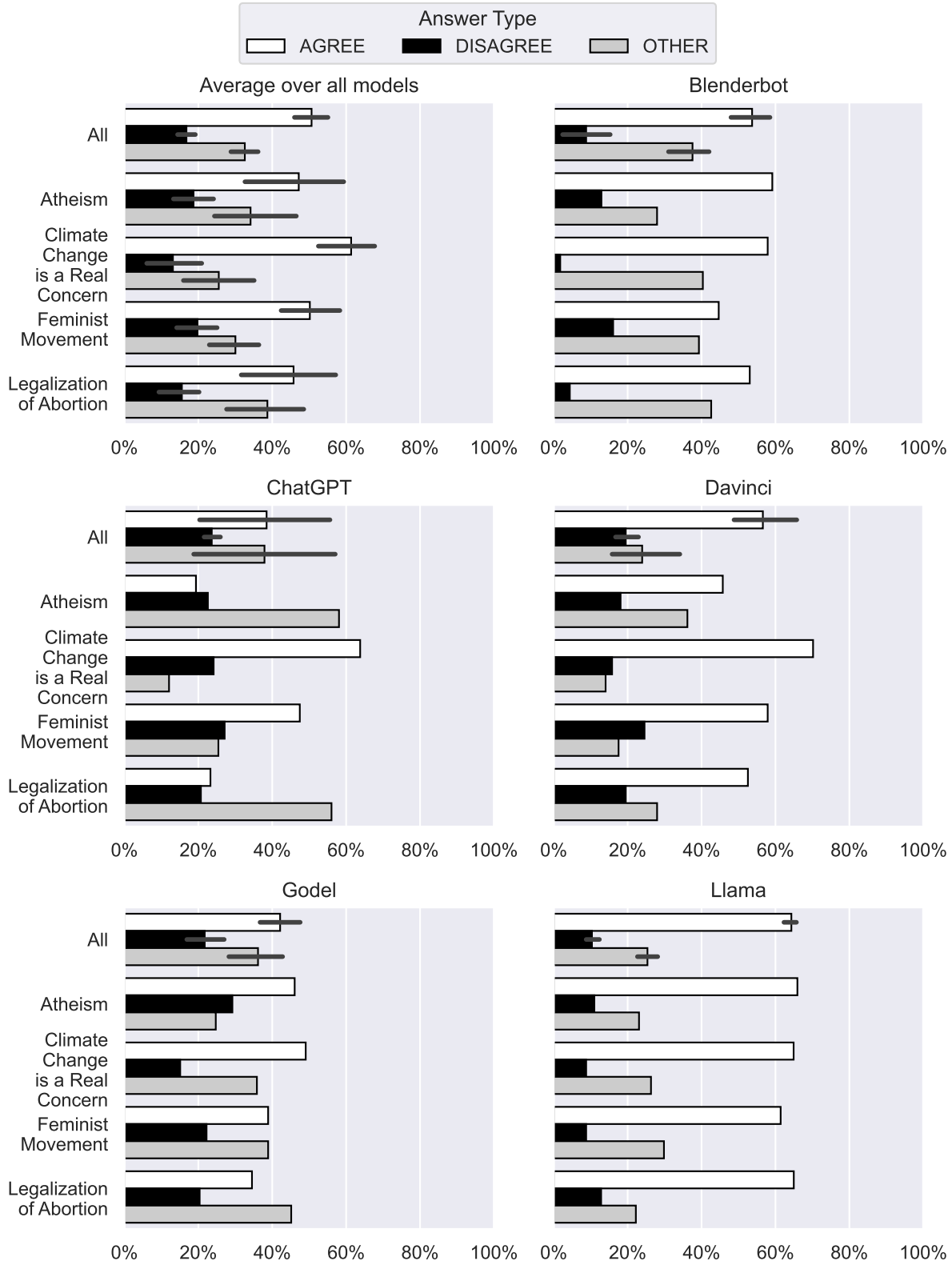

Figure 1: The image shows the relative frequency of the answer categories with one chart for each model and one chart averaged across all models. Each of the charts shows the topic on the vertical axis and the relative amount of `AGREE`, `DISAGREE` and `OTHER` answers as a bar chart. Whenever the chart shows an average value it shows the standard deviation as an error bar as well. The horizontal axis of each chart shows the relative frequency.

Other if none of the above values can be assigned. The (dis)agreement might be expressed implicitly, i.e., a clear statement such as *I agree, you are right* might be missing. The last category occurs mainly in cases (a) when the response presents a rather neutral, balanced answer, (b) when the answer is only a further inquiry of the input or (c) when the answer is rather incomprehensible or not related to the input. Table 2 shows examples of data annotated with labels regarding the (dis)agreement.

Two annotators labeled the data, and a third one resolved disagreements. We provided guidelines and examples to the annotators. We assessed labeling quality using Cohen's Kappa, resulting in an overall moderate inter-annotator agreement with k = 0.49, fair for Llama and Godel with k=0.33 and k=0.37 respectively, moderate for Blenderbot and ChatGPT with k=0.54 each, and substantial for Davinci with k=0.61.

Many annotation issues arose from distinguishing between Other and Agree or Disagree classes. Label confusion frequently occurred with longer and somewhat incoherent chatbot answers (e.g., Llama or Godel answers) or when there was a disparity in opinion strength between the input and the answer.

3. Results

Figure 1 shows the main results of our study. The models we investigated are, on average, 3.02 times more likely to agree with the user than disagree. This effect differs across models and is large for Blenderbot (6.12) and LLama (6.22) and smaller for ChatGPT (1.63), Davinci (2.9) and Godel (1.94). Across all models and topics AGREE is the prevalent annotation except for ChatGPT, which has a very high rate of OTHER for Atheism and Legalization of Abortion.

Figure 2 shows the agreement grouped by stance. The answers are more likely to agree with the stance of the input text when the text is in favor of the topic (34.47%) compared to when the input text is against the topic (15.14%).

4. Discussion

Discussion of LLM answers We explain the high amount of OTHER in the answers of ChatGPT with its tendency to respond neutrally, such as "As a large language model, I do not have an opinion about the topic". We assume that the creators of ChatGPT generated training data for Instruction Based Fine Tuning (Ouyang et al., 2022) to promote this answering style. The results for Blenderbot and Godel also show a high amount of OTHER. These older, less powerful models tend to give

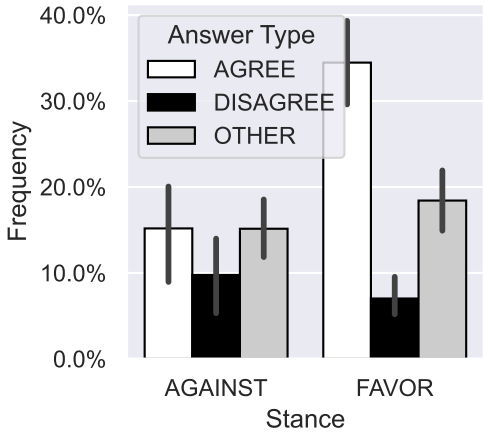

Figure 2: Relative frequency of answer types grouped by stance averaged over all models.

nonsensical answers annotated with OTHER. Also, we found many cases in which ChatGPT states first that it does not have an opinion and utters an opinion afterward, but we did not systematically investigate this behavior.

The models use different answering styles. Blenderbots answers often start with "I agree with you" or "You are right" and then present a short, concise statement about the topic. Llama usually repeats the input and then presents a long answer that continues the first sentence. These long answers tend to go off-topic and are difficult to annotate because the long texts are unclear.

Technical explanation We explain the high level of agreement with the training objective of LLMs: finding the most probable continuation of a text. The training data of LLms are texts from the internet that are predominantly coherent: A text about atheism is more likely to be either in favor of or against atheism than being neutral and changing its opinion from sentence to sentence. In most texts, a sentence is much more likely to agree with the preceding sentence than to disagree, and we can see the same behavior in our study.

Opinionated LLMs Previous work (Hartmann et al., 2023; Santurkar et al., 2023) suggested that LLMs follow a pro-environmental, left-libertarian ideology. We add to this research that the opinions of LLMs are also heavily influenced by their prompts and how one asks about their opinion.

5. Conclusion

Our experiment showed that LLMs tend to agree with their input texts. Therefore, chatbots based on LLMs exhibit similar issues as the echo chamber problem.

6. Ethical considerations

The data from ChatGPT indicates that OpenAI influenced the answer behavior of its chatbots in some topics (atheism and legalization of abortion) and did not influence it in others (climate change is a real concern, feminist movement). If this is the case, we believe that, due to the expected importance of LLM-based chatbots, it is crucial that LLM creators make transparent in how they influenced the answering styles.

Before the launch of the ChatGPT interface to GPT-2.5 by OpenAI in November of 2022, the consequences of the behavior of large language models were relevant mainly to academic researchers. Since then, hundreds of millions of people have presumably interacted with LLMs (Carr, 2023). This number will most likely continue to increase rapidly with the integration of LLMs into ever more products with a combined user base of billions of people, including ChatGPT into Microsoft's Bing search engine (Lardinois, 2023), Google's Bard into Google services like Gmail, Google Docs, and Youtube (Grant, 2023), and Bing Chat, also based on ChatGPT, into Meta's WhatsApp, Messenger, and Instagram (Mehdi, 2023), to name just the best-known services. This development calls for more research into the behavior of large language models, as these tools will mediate more and more of the information people receive and process.

The view that size matters is supported by the fact that the European Union, via the Digital Services Act (DSA), has required so-called very large online platforms (VLOPS) and very large online search engines (VLOSE) to fulfill a range of requirements in order to assess and mitigate "systemic risks" that may emanate from these services, including to freedom of expression and information, to non-discrimination, to a high level of consumer protection and any actual or foreseeable negative effects on civic discourse and electoral processes. The European Commission (2023) has designated Bing, Google Search, Instagram, and YouTube as either VLOPS or VLOES.

In addition, at the time of writing, negotiations continue on the EU level about the "Proposal for a regulation of the European Parliament and of the Council on harmonized rules on Artificial Intelligence (Artificial Intelligence Act)". This includes proposed rules for large language models by both the Council of the European Union and the European Parliament. Among other obligations, The European Parliament (2023) would like to require providers of foundation models (which would include, among others, LLMs) to "demonstrate through appropriate design, testing, and analysis that the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development with appropriate methods such as with the involvement of independent experts, as well as the documentation of remaining non-mitigable risks after development." This view was echoed by a group of academic researchers and civil society representatives who urged the EU to oblige providers and deplers of general purpose AI systems to conduct risk assessment (Aszódi et al., 2023). Albeit unclear what the parties will agree on as the final text, this shows clearly that the lawmaker is concerned with the effects large language models may have, including on the formation of public opinion.

7. Bibliographical References

Nikolett Aszódi, Bettina Berendt, Ian Brown, Nick Diakopoulos, Tim de Jonge, Christina Elmer, Natali Helberger, Clara Helming, Karolina Iwańska, Paul Keller, Frauke Kreuter, Laurens Naudts, Daniel Oberski, Liliane Obrecht, Angela Müller, Estelle Pannatier, Stanislaw Piasecki, João Quintais, Matthias Spielkamp, Alex Tarkowski, Ot van Daalen, Kilian Vieth-Ditlmann, Sophie Weerts, and Frederik Zuiderveen Borgesius. 2023. The ai act and general purpose ai: Charting a path forward. Published: 09/14/2023, Accessed: 10/18/2023.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A. Beam, Marc P. Hauer, Lucien Heitz, Pascal Jürgens, Christian Katzenbach, Benjamin Kille, Beate Klimkiewicz, Wiebke Loosen, Judith Moeller, Goran Radanovic, Guy Shani, Nava Tintarev, Suzanne Tolmeijer, Wouter van Atteveldt, Sanne Vrijenhoek, and Theresa Zueger. 2021. Diversity in News Recommendation. *Dagstuhl Manifestos*, 9(1):43–61. Number: 1 Publisher: Schloss Dagstuhl.

Bill Bishop. 2008. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*, none edition. Houghton Mifflin Harcourt.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020c. Language models are few-shot learners.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *KDD*, pages 535–541. ACM.

David F. Carr. 2023. ChatGPT tops 25 million daily visits. Published: 02/23/2023, Updated: 06/21/2023, Accessed: 10/10/2023.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872, New York, NY, USA. Association for Computing Machinery.

Richard Fletcher and Rasmus Kleis Nielsen. 2018. Are people incidentally exposed to news on social media? A comparative analysis. *New Media & Society*, 20(7):2450–2468. Publisher: SAGE Publications.

Nico Grant. 2023. Google connects a.i. chatbot bard to youtube, gmail and more facts. Published: 09/19/2023, Accessed: 10/10/2023.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv 1503.02531*.

Pascal Jürgens and Birgit Stark. 2022. Mapping Exposure Diversity: The Divergent Effects of Algorithmic Curation on News Consumption. *Journal of Communication*, 72(3):322–344. Publisher: Oxford University Press (OUP).

Frederic Lardinois. 2023. Microsoft launches the new bing, with ChatGPT built in. Published: 02/07/2023, Accessed: 10/10/2023.

Yusuf Mehdi. 2023. Expanding our ai partnership with meta. Accessed: 10/10/2023.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952, Portorož, Slovenia. European Language Resources Association (ELRA).

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings*

of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5356–5371, Online. Association for Computational Linguistics.

Charles E. Osgood and Percy H. Tannenbaum. 1955. The principle of congruity in the prediction of attitude change. *Psychological Review*, 62(1):42–55. Place: US Publisher: American Psychological Association.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Baolin Peng, Michel Galley, Pengcheng He, Chris Brockett, Lars Liden, Elnaz Nouri, Zhou Yu, Bill Dolan, and Jianfeng Gao. 2022. GODEL: Large-scale pre-training for goal-directed dialog.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, Shengjia Zhao, Arun Vijayvergiya, Eric Sigler, Adam Perelman, Chelsea Voss, Mike Heaton, Joel Parish, Dave Cummings, Rajeev Nayak, Valerie Balcom, David Schnurr, Tomer Kaftan, Chris Hallacy, Nicholas Turley, Noah Deutsch, Vik Goel, Jonathan Ward, Aris Konstantinidis, Wojciech Zaremba, Long Ouyang, Leonard Bogdonoff, Joshua Gross, David Medina, Sarah Yoo, Teddy Lee, Ryan Lowe, Dan Mossing, Joost Huizinga, Roger Jiang, Carroll Wainwright, Diogo Almeida, Steph Lin, Marvin Zhang, Kai Xiao, Katarina Slama, Steven Bills, Alex Gray, Jan Leike, Jakub Pachocki, Phil Tillet, Shantanu Jain, Greg Brockman, Nick Ryder, Alex Paino, Qiming Yuan, Clemens Winter, Ben Wang, Mo Bavarian, Igor Babuschkin, Szymon Sidor, Ingmar Kanitscheider, Mikhail Pavlov, Matthias Plappert, Nik Tezak, Heewoo Jun, William Zhuk, Vitchyr Pong, Lukasz Kaiser, Jerry Tworek, Andrew Carr, Lilian Weng, Sandhini Agarwal, Karl Cobbe, Vineet Kosaraju, Alethea Power, Stanislas Polu, Jesse Han, Raul Puri, Shawn Jain, Benjamin Chess, Christian Gibson, Oleg Boiko, Emy Parparita, Amin Tootoonchian, Kyle Kosic, and Christopher Hesse. 2023. Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed: 2023-09-20.

Daniel Stegmann, Melanie Magin, and Birgit Stark. 2022. Echo chambers. In Andrea Ceron, editor, *Encyclopedia of Technology Politics*, pages 210–216. Edward Elgar Publishing LTD.

Cass. R. Sunstein. 2001. *Republic.com*. Princeton Univ. Press, Princeton, NJ.

The European Commission. 2023. Digital services act: Commission designates first set of very large online platforms and search engines. Published: 04/25/2023, Accessed: 10/10/2023.

The European Parliament. 2023. Draft: Proposal for a regulation of the european parliament and of the council on harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. Published: 05/16/2023, Accessed: 10/10/2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.